Genome-wide trans-ethnic meta-analysis identifies seven genetic loci influencing erythrocyte traits and a novel role for *RBPMS* in erythropoiesis

Running title: Trans-ethnic Erythrocyte GWAS

Authors:

Frank JA van Rooij,[1] Rehan Qayyum,[2] Albert V Smith,[3,4] Yi Zhou,[5,6] Stella Trompet,[7,8] Toshiko Tanaka,[9] Margaux F Keller,[10] Li-Ching Chang,[11] Helena Schmidt,[12] Min-Lee Yang,[13] Ming-Huei Chen,[14,15] James Hayes,[16] Andrew D Johnson,[15] Lisa R Yanek,[2] Christian Mueller,[17] Leslie Lange,[18] James S Floyd,[19] Mohsen Ghanbari,[1] Alan B Zonderman,[20] J Wouter Jukema,[7] Albert Hofman,[1,21] Cornelia M van Duijn,[1] Karl C Desch,[22] Yasaman Saba,[12] Ayse B Ozel,[23] Beverly M Snively,[24] Jer-Yuarn Wu,[11,25] Reinhold Schmidt,[26] Myriam Fornage,[27] Robert J Klein,[16] Caroline S Fox,[15] Koichi Matsuda,[28] Naoyuki Kamatani,[29] Philipp S Wild,[30,31,32] David J Stott,[33] Ian Ford,[34] P Eline Slagboom,[35] Jaden Yang,[36] Audrey Y Chu,[37] Amy J Lambert,[38] André G Uitterlinden,[1,39] Oscar H Franco,[1] Edith Hofer,[26,40] David Ginsburg,[23] Bella Hu,[5,6] Brendan Keating,[41,42] Ursula M Schick,[43,44] Jennifer A Brody,[19] Jun Z Li,[23] Zhao Chen,[45] Tanja Zeller,[17,46] Jack M Guralnik,[47] Daniel I Chasman,[48,37] Luanne L Peters,[38] Michiaki Kubo,[49] Diane M Becker,[2] Jin Li,[50] Gudny Eiriksdottir,[4] Jerome I Rotter,[51] Daniel Levy,[15] Vera Grossmann,[52] Chien-Hsiun Chen,[11,25] The BioBank Japan Project,[53] Paul M Ridker,[54,37] Hua Tang,[55] Lenore J Launer,[56] Kenneth M Rice,[57] Ruifang Li-Gao,[58] Luigi Ferrucci,[9] Michelle K Evans,[59] Avik Choudhuri,[5,60] Eirini Trompouki,[61,62] Brian J Abraham,[63] Song Yang,[5,6] Atsushi Takahashi,[29] Yoichiro Kamatani,[29] Charles Kooperberg,[64,65] Tamara B Harris,[56] Sun Ha Jee,[66] Josef Coresh,[67] Fuu-Jen Tsai,[25] Dan L Longo,[68] Yuan-Tsong Chen,[11] Janine F Felix,[1] Qiong Yang,[69,15] Bruce M Psaty,[70,71] Eric Boerwinkle,[72] Lewis C Becker,[2] Dennis O Mook-Kanamori,[73,58,74] James G Wilson,[75] Vilmundur Gudnason,[3,4] Christopher J O'Donnell,[15] Abbas Dehghan,[1] L. Adrienne Cupples,[69,15] Michael A Nalls,[10] Andrew P Morris,[76,77] Yukinori Okada,[78,29] Alexander P Reiner,[79,80] Leonard I Zon,[5,6] Santhi K Ganesh,[13 *]

Affiliations:

[1]Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands. [2]GeneSTAR Research Program, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [3]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [4]Icelandic Heart Association, Kopavogur, Iceland. [5]Harvard Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [6]Stem Cell Program and Division of Hematology/Oncology, Children's Hospital Boston, Pediatric Hematology/Oncology at DFCI, Harvard Stem Cell Institute, Harvard Medical School and Howard Hughes Medical Institute, Boston, MA 02115, USA. [7]Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands. [8]Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands. [9]National Institute on Aging, National Institutes of Health, Baltimore, MD USA. [10]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Besthesda, MD USA 20892. [11]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. [12]Institute of Molecular Biology and Biochemistry, Centre for Molecular Medicine, Medical University of Graz. [13]Division of Cardiovascular Medicine, Department of Internal Medicine, Department of Human Genetics, University of Michigan, 1500 E. Medical Center Drive Ann Arbor, MI 48109. [14]Department of Neurology ,Boston University School of Medicine. [15]Framingham Heart Study, Population Sciences Branch, Division of Intramural Research National Heart Lung and Blood Institute, National Institutes of Health. [16]Icahn Institute for Multiscale Biology, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029. [17]Department of General and Interventional Cardiology, University Heart Centre Hamburg-Eppendorf, Hamburg, Germany. [18]Department of Genetics, University of North Carolina, Chapel Hill, NC, 27599 USA. [19]Department of Medicine, University of Washington, Seattle, WA. [20]National Institute on Aging, National Institutes of Health, Bethesda, MD, USA. [21]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Mass, USA. [22]Department of Pediatrics and Communicable Disease, University of Michigan, Ann Arbor, MI 48109.. [23]Department of Internal Medicine, Human Genetics, Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, MI 48109.. [24]Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States of America. [25]School of Chinese Medicine, China Medical University, Taichung, Taiwan. [26]Clinical Division of Neurogeriatrics, Department of Neurology, Medical University Graz, Austria. [27]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA. [28]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan. [29]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan. [30]Center for Thrombosis and Hemostasis (CTH), University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany. [31]German Center for Cardiovascular Research (DZHK), Partner Site RhineMain, Mainz, Germany. [32]Preventive Cardiology and Preventive Medicine, Center for Cardiology, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany. [33]Institute of Cardiovascular and Medical Sciences, Faculty of Medicine, University of Glasgow, United Kingdom. [34]Robertson Center for Biostatistics, University of Glasgow, United Kingdom. [35]Department of Medical Statistics and Bioinformatics, Section of Molecular Epidemiology, Leiden University Medical Center, Leiden , The Netherlands. [36]Quantitative Sciences Unit, School of Medicine, Stanford University. [37]Division of Preventive Medicine, Brigham and Women's Hospital and Harvard

Medical School, Boston MA 02215 USA. [38]The Jackson Laboratory, Bar Harbor, Maine, USA. [39]Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. [40]Institute of Medical Informatics, Statistics and Documentation, Medical University Graz, Austria. [41]Center for Applied Genomics, Children's Hospital of Philadelphia, PA, USA. [42]Dept of Pediatrics, University of Pennsylvania, PA, USA. [43]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [44]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [45]Department of epidemiology and biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona. [46]German Center for Cardiovascular Research (DZHK), Partner Site Hamburg, Lübeck, Kiel, Hamburg, Germany. [47]Department of Epidemiology and Public Health, University of Maryland School of Medicine. [48]Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston MA 02115 USA. [49]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan. [50]Cardiovascular Medicine Division, Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94304. [51]Institute for Translational Genomics and Population Sciences, Departments of Pediatrics and Medicine, LABioMed at Harbor-UCLA Medical Center, Torrance, CA USA. [52]Center for Thrombosis and Hemostasis (CTH), University Medical Center Mainz, Mainz, Germany. [53]The BioBank Japan Project, Japan. [54]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston MA 02115 USA. [55]Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA. [56]Laboratory of Epidemiology, Demography, and Biometry, National Institute on Aging, Intramural Research Program, National Institutes of Health, Bethesda, Maryland, USA. [57]Department of Biostatistics University of Washington, Seattle, WA. [58]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. [59]Health Disparities Research Section, Clinical Research Branch, National Institute on Aging, National Institutes of Health, Baltimore, Maryland, United States of America. [60]Stem Cell Program and Division of Hematology/Oncology, Children's Hospital Boston, Pediatric Hematology/Oncology at DFCI, Harvard Stem Cell Institute, Harvard Medical School and Howard Hughes Medical Institute, Boston, MA 02115, USA. [61]Max Planck Institute of Immunobiology and Epigenetics, Freiburg 79108, Germany. [62]Stem Cell Program and Division of Hematology/Oncology, Children's Hospital Boston, Pediatric Hematology/Oncology at DFCI, Harvard Stem Cell Institute, Harvard Medical School and Howard Hughes Medical Institute, Boston, MA 02115, USA. [63]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. [64]Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, WA. [65]Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA. [66]Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea. [67]Johns Hopking Bloomberg School of Public Health, George W. Comstock Center for Public Health Research and Prevention, Comstock Center & Cardiovascular Epidemiology, Welch Center for Prevention, Epidemiology and Clinical Research. [68]Clinical Research Branch, National Institute on Aging, Baltimore, Maryland, United States of America. [69]Department of Biostastics, Boston University of Public Health. [70]Departments of Epidemiology, Health Services, and Medicine, University of Washington, Seattle, WA. [71]Group Health Research Institute, Group Health Cooperative, Seattle, WA. [72]Human Genetics Center 1200 Herman Pressler E-447, Houston, TX 77030. [73]Department of BESC, Epidemiology Section, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia. [74]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands. [75]Department of Physiology and Biophysics, University of Mississippi Medical Center,

Jackson, MS, 39216 USA. [76]Department of Biostatistics, University of Liverpool, Block F, Waterhouse Building, 1-5 Brownlow Street, Liverpool L69 3GL, UK. [77]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [78]Department of Human Genetics and Disease Diversity, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo 113-0085, Japan. [79]Department of Epidemiology, University of Washington, Seattle, Washington, United States of America. [80]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America.

* Correspondence: sganesh@umich.edu

Corresponding Author:
Santhi K. Ganesh
Division of Cardiovascular Medicine, Department of Internal Medicine
Department of Human Genetics
University of Michigan
1500 E. Medical Center Drive Ann Arbor, MI 48109
Tel (734)764-4500| Fax (734)936-8266
sganesh@umich.edu

Word count Abstract : 149

Word count Main Text : 3471

Abstract

Genome-wide association studies (GWAS) have identified loci for erythrocyte traits in primarily European ancestry populations. We conducted GWAS meta-analyses of six erythrocyte traits in 71,638 individuals from European, East-Asian, and African ancestries using a Bayesian approach to account for heterogeneity in allelic effects and variation in the structure of linkage disequilibrium between ethnicities. We identified seven novel loci for erythrocyte traits including a novel locus (*RBPMS/GTF2E2*), associated with mean corpuscular hemoglobin and mean corpuscular volume. Statistical fine-mapping at this locus pointed to the *RBPMS* gene at this locus and excluded the nearby *GTF2E2* gene. Using zebrafish morpholino to evaluate loss-of-function, we observed a strong *in vivo* erythropoietic effect for *RBPMS* but not for *GTF2E2*, supporting the statistical fine-mapping at this locus, and demonstrating that *RBPMS* is a novel regulator of erythropoiesis. Our findings show the utility of trans-ethnic GWAS for discovery and characterization of genetic loci influencing hematologic traits.

Introduction

Erythrocytes disorders are common world-wide, contributing to substantial morbidity and mortality.[1] Erythrocyte counts and indices are heritable (estimated h^2 = 0.40-0.90 [2–4]), exhibit different patterns across ethnic groups, and have been influenced by selection in various ethnic groups, most notably for protection against infection against parasites such as those that cause malaria.[5–7] Erythrocyte traits have been studied most extensively in European ancestry populations,[8–10] and smaller studies in non-European populations have shown both shared and distinct genetic loci influencing erythrocyte traits.[11,12]

Trans-ethnic meta-analysis of genome-wide association studies (GWAS) data offers improved signal detection in a combined meta-analysis when heterogeneity of allelic effects, allele frequencies and differences in linkage disequilibrium (LD) between ethnicities are accounted for. Trans-ethnic meta-analysis can also enable fine-mapping of association intervals by evaluating differences in LD structure between diverse populations, thereby enhancing the detection of causal variants.[13]

We conducted trans-ethnic GWAS meta-analyses with the goal of elucidating the genetic architecture of erythrocyte traits, to evaluate whether: (i) combining data across populations of diverse ancestry may improve power to detect associations for erythrocyte traits; and (ii) differences in LD structure can be exploited to identify causal variants driving the observed associations with common SNPs. In this study, we analyzed GWAS summary statistics from 72,630 individuals from three diverse populations of European (EUR), East Asian (EAS), and African (AFR) ancestry. We conducted replication analyses in independent samples and performed functional testing to support our approach to fine-mapping.

Subjects and Methods

Study samples

We aggregated HapMap-imputed GWAS results from 71,638 individuals represented in several cohorts embedded in the CHARGE Consortium (40,258 individuals of EUR ancestry), the RIKEN / BioBank Japan Project and AGEN cohorts (15,252 individuals of EAS ancestry), and the COGENT Consortium (16,128 individuals of AFR ancestry). Phenotypic information on all participating cohorts is provided in **Table S1** and has been reported previously.[8,11,12,14,15] We conducted replication analyses of novel trait-loci associations in six independent studies: the Gutenberg Health Study (GHS cohorts 1 and 2, both EUR ancestry), the Genes and Blood-Clotting Study (GBC, EUR ancestry), the NEO study (EUR ancestry) , the JUPITER trial (EUR ancestry), and the HANDLS study (AFR ancestry) [16–21] (total replication size N= 16,389).

Erythrocyte phenotype modelling

We analyzed six erythrocyte traits; hemoglobin concentration (Hb, g/dL), hematocrit (Hct, percentage), mean corpuscular hemoglobin (MCH, picograms), mean corpuscular hemoglobin concentration (MCHC, g/dL), mean corpuscular volume (MCV, femtoliters), and red blood cell count (RBC, 1M cells/cm^3). Trait units were harmonized across all studies. MCH, MCHC, MCV and RBC were transformed to obtain normal distributions. We excluded samples deviating more than 3 SD from the ethnic and trait specific mean within each contributing study, because we focused on determinants of variation in the general population rather than on specific hematological diseases which are overrepresented at the extremes of the trait distribution (**Table S2**).

Statistical analyses

For the initial GWA analyses, cohorts used linear regression or linear mixed models to assess the association of the SNPs meeting the quality control criteria with each of the six traits. An additive genetic model was used and the regressions were adjusted for age, sex and study site, if applicable (**Appendix**).

For the ethnic-specific meta-analyses, GWAS results of SNPs with a minor allele frequency (MAF) ≥1% were analyzed in a fixed-effect meta-analysis (METAL software[22]) within each ancestry group, with genomic control (GC) correction of the individual GWAS results of each contributing cohort and the final meta-analysis results.[23]

For the trans-ethnic meta-analyses, the three sets of the ethnic-specific meta-analysis summary statistics were then combined with three approaches. First, fixed effects meta-analyses of the three sets of ethnic-specific summary results were conducted, with GC correction of the final trans-ethnic meta-analysis results. Secondly, the three ethnic-specific data sets were combined using MANTRA (Meta-Analysis of Trans-ethnic Association studies),[24] which uses a Bayesian algorithm, allowing for heterogeneity in allelic effects arising as a result of variation in LD structure across different ancestry groups. Finally, the three sets of ethnic-specific results were analyzed by means of the Han and Eskin RE2 model, a meta-analysis method developed for higher statistical power under heterogeneity.[25]

For the replication analyses, the independent replication cohorts provided linear regression results for the novel trait-locus combinations. Their results were meta-analyzed with a fixed effects inverse variance weighted method (METAL) and the RE2 methodology. Additionally, we meta-analyzed their results with the discovery data using fixed-effects, MANTRA, and RE2 methods.

We used the MANTRA results to fine-map the regions of trait-associated index SNPs. We defined regions by identifying variants within a 1 Mb window around each index SNP (500kb upstream and 500 kb downstream). For each SNP in a region, the posterior probability that this SNP is driving the region's association signal was calculated by dividing the SNP's BF by the summation of the BFs of all SNPs in the region. Credible sets (CS) were subsequently created by sorting the SNPs in each region in descending order based on their BF (so starting with the index SNP since this SNP has the region's largest BF by definition). Going down the sorted list, the SNPs' posterior probabilities were summed until the cumulative value exceeded 99% of the total cumulative posterior probability for all SNPs in the region. The length of a CS was expressed in base pairs. We compared 99% credible sets for the trans-ethnic results and the results of a EUR-only MANTRA analysis.[13,24,26] For the MANTRA fine-mapping analyses, a less stringent threshold value of log10 BF > 5 was applied, because we wanted to include previously identified regions which may not have showed up in the more stringent MANTRA discovery analyses.

ENCODE annotation

We evaluated the SNPs identified in the discovery analyses against the ENCODE Project Consortium's database of functional elements in the K562 erythroleukemic line.[27]

Experiments in zebrafish

To substantiate the fine-mapping of the *RBPMS*/*GTF2E2* region biologically, we tested the effect of morpholino knockdown in zebrafish for both *RBPMS* and *GTF2E2* orthologous genes, followed by assays of erythrocyte development. For the selected genes, gene synteny analysis was performed and peptide homology comparison and domain structure were used when no syntenic region was previously annotated. For each gene, morpholino constructs were designed incorporating information about gene structure and translational initiation sites (Gene-Tool Inc., Philomath, OR). MOs targeting each transcript, were injected into single-cell embryos at 1, 3, and 5 ng/embyo doses to find an optimal dose at which there were

minimum non-specific toxicity. Post-injection, embryos were collected at specified time points, 16-18 ss, 22-26 hpf, and 48 hpf using both standard morphological features of the whole embryo and hours post-fertilization (hpf) to stage to minimize differences in embryonic development staging caused by the MO injection.[28,29] The embryos were then assayed for hematopoietic development by whole-mount in situ hybridization and benzidine staining. For the globin transcription, developing erythrocytes in the intermediate cell mass of the embryos were assayed by embryonic β-globin 3 expression at the 16 somite stage, or 16-18 hpf.[29] Further detailed methods are provided in the **Appendix**.

Chromatin Immunoprecipitation (ChIP) and Assay for Transposase Accessible Chromatin (ATACseq) in human CD34+ cell lines

CD34+ cells were expanded and differentiated according to the detailed methods provided in the Appendix. For ChIP-seq experiments in human CD34+ cells, Gata1 (Santa Cruz sc265X), Gata2 (Santa Cruz sc9008X) and H3K27ac (Abcam ab4729) antibodies were used as described previously. ChIP-Seq and ATACseq reads were aligned to the human reference genome (hg19) and high-confidence peaks of ChIP-Seq signal were identified and analyzed. A q-value threshold of enrichment of 0.05 was used for all datasets. Detailed methods are provided in the **Appendix**.

Evaluation in mouse crosses

To further affirm the novel trait-loci we identified, and in an attempt to further fine-map the intervals identified in our discovery analyses through cross-species comparisons, we evaluated the new loci in syntenic regions in twelve inter-strain mouse QTL crosses.[30] The methods for the original mouse cross have been previously described.[30] Mice from 12 different strains were inter-crossed and the same erythrocyte traits we have studied by GWAS were measured in peripheral blood (**Appendix**).

Results

We identified 44 previously reported loci [7–12,31–35] (**Table S3**) and nine novel significant trait-locus associations at seven loci ($P < 5x10^{-8}$ or $\log_{10}BF > 6.1$, **Table 1**). *SHROOM3* was simultaneously identified in an exome chip analysis by our group in overlapping samples.[36] Ethnic-specific results are presented in **Table S4**. Regional association plots are shown for each region in **Figure S1**, showing ethnic-specific results, the trans-ethnic meta-analysis and plots of pairwise LD across the regions for EUR, EAS and AFR ancestry.

Five of the novel trait-loci showed a significant association in the fixed-effects trans-ethnic METAL analyses, in the Bayesian MANTRA analyses, and in the RE2 analyses; these were *TMEM163*/*ACMSD* for Hct, *PLCL2:rs2060597* for MCH, and *ID2*, *PLCL2:rs9821630*, and *RBPMS* for MCV. Two loci (*MET* and *FOXS1*) showed a borderline significant effect in METAL and RE2, and a strong significant effect in MANTRA for HB and MCV respectively. The association of rs2979489 (*RBPMS*) further showed a strong association with MCH in the multi-ethnic Bayesian meta-analysis and in the RE2 model, but was not detected in the multi-ethnic fixed-effects meta-analysis, nor in any of the ethnic-specific meta-analyses for this trait. Interestingly, MCH and MCV are correlated traits, yet strong heterogeneity of effect was observed for this SNP's association with MCH only, as indicated by both METAL (I^2 statistic 94%, *P* value Cochran's Q statistic of heterogeneity 6.48×10^{-8}) and MANTRA (posterior probability of heterogeneity = 1) (**Table 1**). Inspection of the discovery data sets showed that one of the African-American cohorts supplied data for MCV but not for MCH, which resulted in a stronger positive association of rs2979489 with MCH than with MCV in the AFR meta-analyses. This phenomenon was accompanied by greater evidence of heterogeneity for MCH in the trans-ethnic meta-analyses because the EUR and EAS associations were in the opposite direction to that observed in the AFR meta-analysis. The MANTRA and RE2 analyses were able to account for this heterogeneity, and thus yield a stronger result as compared to METAL for this trait-locus.

Replication analyses

In the meta-analyses of the replication cohorts the trait-SNP combinations HT-*TMEM163*/*ACMSD* and MCH-*RBPMS* achieved a Bonferroni-corrected significance threshold with both fixed effects and RE2 methods ($p<0.05/9$). *ID2* was Bonferroni-significant in the fixed-effects model and nominally significant in the RE2 model. Furthermore, we found nominal significance for MCV-*RBPMS* (fixed-effects analyses) and *FOXS1* (fixed-effects and RE2) (**Table S5**).

When we compared the discovery and replication combined meta-analyses with the discovery analyses alone, we observed stronger associations for Hct-*TMEM163/ACMSD*, MCH-*PLCL2*, MCV-*ID2*, and MCV-*RBPMS* in all three models (fixed-effects, MANTRA and RE2). For MCH-*RBPMS*, we found a stronger association in the fixed-effects analysis.(**Table S6)**.

Statistical finemapping

We found that 31 trait-specific trans-ethnic 99% CSs showed a decrease in length of at least 50% as compared to their EUR only CS counterparts (26 unique loci across the six erythrocyte traits) (**Table S7**).

Among the novel loci identified in this study, the chromosome 8 *RBPMS* locus showed fine-mapping according to this criterion (**Table 2**, **Figure 1**). For MCH, the EUR credible set spanned 204,200bp, encompassing the *RBPMS* and *GTF2E2* genes. The multi-ethnic credible set comprised just one SNP, rs2979489, within the first intron of the *RBPMS* gene (**Figure 1**). Remarkably, this associated SNP rs2979489 is located adjacent to a GATA-motif where a gradual switch of binding from GATA2 to GATA1 takes place during commitment of human CD34 progenitors towards erythroid lineage (**Figure 2**, Bottom left Panel). Moreover, an assay for chromatin accessibility sites (ATAC-seq) and H3K27a ChIP-seq clearly identify that the genomic region proximal to this SNP is actively regulated during human erythroid differentiation **(Figure 2**, Bottom right Panel).

Among the known loci, fine mapping narrowed signals as shown in **Table S7**. Interestingly, trans-ethnic fine-mapping of the *XRN1* locus (MCH) led us to the rs6791816 polymorphism. Van der Harst *et al* also identified the same SNP in their exploration of nucleosome-depleted regions (NDRs, representing active regulatory elements for erythropoeisis) in a follow-up analysis of their GWAS results.[10] By means of subsequent Formaldehyde-Assisted Isolation of Regulatory Elements followed by next-generation sequencing (FAIRE-seq), they pinpointed rs6791816 as an NDR SNP in LD with their initial index SNP for MCH and MCV.

Furthermore, fine-mapping of both the *MPND* locus (MCH) and *SH3GL1* locus (MCV) pointed to the rs8887 SNP within the 3'UTR of *PLIN4*. The rs8887 SNP minor allele has been shown experimentally to create a novel seed site for miR-522, resulting in decreased PLIN4 expression.[37] Furthermore, miR-522 is an expressed miRNA in circulating blood.[38] These data suggest that an allele-specific miR-522 regulation of PLIN4 by rs8887 could serve as a functional mechanism underlying the identified association.

We additionally showed fine mapping in several other intervals (**Table S7**) with fine-mapped genes about which less is known about their potential biologic role in erythropoeisis or red blood cell function. These regions are of interest for further hypothesis generation based upon the GWAS findings.

ENCODE analyses

We further evaluated the SNPs from the chromosome 8 *RBPMS* region against the ENCODE Project Consortium's database of numerous functional elements in the K562 erythroleukemic line.[27] The lone SNP that was fine mapped at the locus, rs2979489, was found in a strong enhancer element as defined by Segway, supporting a functional role for this SNP and the *RBPMS* gene. The other SNPs in the *RBPMS* region, excluded by the statistical fine-mapping exercise, were not annotated as regulatory in the ENCODE data (**Table S8).**

Experiments in zebrafish

We identified a novel erythropoietic effect for the zebrafish *rbpms* gene. Both embryonic globin expression at 16 somite stage and o-dianisidine/benzidine staining at 48 hours post fertilization significantly decreased in morphants, indicating a decrease in both globin transcription and Hb levels (**Figure 3**). This loss-of-function finding is consistent with a decreased mean erythrocyte Hb content observed in our human association results. In zebrafish, the *rbpms* orthology mapping included *rbpms2a, rbpms2b* and *rbpms*, and loss-of-function phenotypes of all orthologs were tested experimentally. The results suggested a clear erythropoietic effect with limited functional compensation of the genes in the rbpms family in zebrafish during embryonic erythropoiesis. On the other hand, morpholino knockdown experiments with the zebrafish ortholog of the *GTF2E2* gene did not show an apparent erythropoietic effect.

Review of the human association results showed no evidence of pleiotropy across the RBPMS family of genes and denote that the human association is specific to *RBPMS* (**Supplemental Data**). This review was conducted because the orthology in the fish led to inclusion of *rbpms2* in the zebrafish analyses as well. These findings indicate that the statistical fine-mapping was useful to home in on *RPBMS* as a causal gene influencing erythropoiesis.

Evaluation in mouse crosses

In the eight novel regions from our discovery analysis, six had evidence of cross-species validation by evidence of syntenic gene within the linkage peak in the mouse QTL results (**Table 3**). However, the human GWAS intervals were not narrowed by the mouse QTL results for any of these loci (**Table S9**).

Discussion

We conducted GWASs and meta-analyses of six erythrocyte traits (Hb, Hct, MCH, MCHC, MCV, and RBC) in 71,638 individuals from European, Asian, and African-American ancestry.

While prior genome-wide association studies have identified loci associated with erythrocyte traits through the analysis of ancestrally homogenous cohorts and consortia, largely biased towards European ancestry studies, trans-ethnic analysis has not previously been performed while accounting for differences in genetic architecture in ethnically diverse groups.

We identified seven novel loci for erythrocyte traits (nine locus-trait combinations) and replicated 44 previously identified loci. We fine-mapped several known and new loci. One fine-mapped locus led us to a region on chromosome 8 associated with MCH and MCV.

In the chromosome 8 *RBPMS/GTF2E2* locus, the index variant rs2979489, which was associated with MCV and MCH and highlighted in the trans-ethnic fine-mapping analyses, is located within the first intron of the *RBPMS* (RNA binding protein with multiple splicing) gene, notably at an open chromatin site at which a switch of GATA1/2 binding occurs during erythroid differentiation. The *RBPMS* protein product regulates a variety of RNA processes, including pre-mRNA splicing, RNA transport, localization, translation, and stability.[39,40] The *RBPMS* gene is expressed at relatively low levels in mammalian erythroblasts and the protein product has not been detected in mature human erythrocytes.[41,42]

The rs2979489 polymorphism showed remarkable high heterogeneity in effect on the MCH trait across the different ethnicities, with different directions of effect for the AFR meta-analysis results compared to the EUR and ASN findings. If the variant is causal, this pattern of association could reflect gene-environment interaction. In this case different exposures in AFR compared to EUR/ASN populations may lead to a marginal effect of the SNP in opposing directions by different selection pressures. If however rs2979489 is not causal, but rather a marker in LD with the causal variant, then the opposing direction of effects could reflect very different LD structures in the different populations, also indicating selection, or theoretically it could even reflect different causal variants in AFR and EUR/EAS - and rs2979489 being just in strong LD with both causal variants.

The SNP rs2979489 is located adjacent to a GATA-motif where a gradual switch of binding from GATA2 to GATA1 takes place during commitment of human CD34 progenitors towards erythroid lineage. These observations suggest that rs2979489 localize at a potential regulatory site where a modulation of erythroid cell differentiation occurs and the presence of rs2979489 may lead to observed red cell trait alterations in human populations, possibly through regulation of *RBPMS* gene expression timing, level, and/or splicing variation. Although *RBPMS* previously had no known role in haematopoiesis or more specifically in erythropoiesis, *RBPMS* has been previously shown to be upregulated in transcriptional profiles of murine and human hematopoietic stem cells.[43–45] Its role may be at much earlier stages during the differentiation of erythrocytes from erythroblasts and/or hematopoietic stem cells. *RBPMS* is known to physically interact with Smad2, Smad3, and Smad4 and stimulate smad-mediated transactivation through enhanced Smad2 and Smad3 phosphorylation and associated promotion of nuclear accumulation of Smad proteins.[46] These SMAD proteins are known to regulate the TGF-β mediated regulation of hematopoietic cell fate and erythroid differentiation.[47] *RBPMS* has four annotated transcript isoforms, and further delineation of the tissue-specificity, timing of expression, and function of these transcripts in the context of the genetic variant we identified warrants further study.

Among the additional six loci, we identified two loci in which the index SNP was located within annotated genes, rs6430549 in *ACMSD* (aminocarboxymuconate semi aldehyde decarboxylase, intronic and rs2299433 in *MET* (mesenchymal epithelial transition factor, intronic). No previous hematologic role has been described for both regions. Variants in the chromosome 2q21.3 *ACMSD* region have previously been associated with blood metabolite levels, obesity and Parkinson's Diseases.[48–50] A genetic variant in the first intron of the *MET* gene was significantly associated with both Hb and Hct; however association was not observed in replication samples possibly due to lower power in the replication experiment. Three additional loci were intergenic but close to a coding gene (rs10929547 near *ID2*

(inhibitor of DNA binding 2, dominant negative helix-loop-helix protein) and rs6121246 near *FOXS1* (forkhead box S1), and rs2060597 approximately 40 kbp upstream of the *PLCL2* (phospholipase C-like 2) gene). The roles of variants in these regions in determining erythrocyte traits are unknown.[41,51]

In the statistical fine-mapping analyses, the trans-ethnic meta-analysis approach resulted in smaller 99% credible intervals in all of the novel loci identified in this study. Since these loci were identified in analyses that accounted for heterogeneity in allelic effects between ethnic groups, in which the heterogeneity may be due to variation in LD patterns, we examined the LD patterns in these loci. Not surprisingly, we noted that the consistent decrease in the size of 99% credible interval across all loci is likely due to the inclusion of cohorts of African ancestry, an ethnic group with generally smaller LD blocks throughout the genome. The loss-of-function screens in zebrafish for the chromosome 8 signal suggested that these analyses successfully identified a single gene (*RBPMS*) with erythropoietic effect within one of the fine-mapped intervals. We also fine-mapped previously known regions such as the chromosome 6p21.1 region associated with RBC count, highlighted *CCND3*, which has been experimentally shown to regulate RBC count experimentally in a knock-out mouse model.[52] These examples suggest that attempts to refine association signals using these types of approaches in existing samples may yield functional candidates for further mechanistic hypothesis testing, which is a major goal of GWAS.

Trans-ethnic genome-wide meta-analyses of common variants have aided in the characterization of genetic loci for various complex traits.[13,53–55] Our data demonstrate the benefits of trans-ethnic genome-wide meta-analysis in identifying and fine-mapping genetic loci of erythrocyte traits. By exploiting the differences in genetic architecture of the associations within these loci in various ethnic groups, we may identify causal genes influencing clinically relevant hematologic traits. Use of a similar approach for other complex

traits is likely to provide deeper insights into the biological mechanisms underlying human traits.

Appendix

Genotyping

In brief, the cohorts comprise unrelated individuals, except for the Framingham Heart Study (related individuals of European ancestry) and GeneSTAR (related individuals of European or African ancestry). SNPs with a minor allele frequency < 1%, missingness >5% or HWE $P < 10^{-7}$ were excluded. Genotypes were imputed to approximately 2.5 million SNPs using HapMap Phase II CEU. The RIKEN and the BioBank Japan Project and AGEN cohorts comprise unrelated individuals of East-asian ancestry (EAS). SNPs with a minor allele frequency < 0.01, missingness >1% or HWE $P < 10^{-7}$ were excluded. Individuals with a call rate < 98% were excluded as well. Genotypes were imputed to approximately 2.5 million SNPs using HapMap Phase II JPT and CHB. The COGENT consortium cohorts comprise individuals of African-American ancestry (AFR). SNPs with a minor allele frequency < 1% or missingness >10% were excluded. Genotypes were imputed to approximately 2.5 million SNPs using HapMap Phase II CEU and YRI.

Cohort specific GWAS

Each cohort used linear regression to assess the association of all SNPs meeting the quality control criteria with each of the six traits separately. An additive genetic model was used and the regressions were adjusted for age, sex and study site (if applicable). The Framingham Heart Study and the GeneSTAR study used linear mixed effects models to account for relatedness, and these models included adjustment for principal components(address stratification comment). Multi-center studies adjusted for study site.

Ethnic-specific GWAS meta-analyses

We performed three ethnic-specific GWAS fixed effects inverse variance-weighted meta-analyses for each trait using METAL software.[22] SNPs with an imputation quality < 30% were excluded and study results were adjusted for genomic inflation factors.[23]

Trans-ethnic meta-analyses

For each trait we performed a trans-ethnic fixed-effect inverse variance-weighted meta-analysis of the EUR, EAS, and AFR GWAS summary statistics using METAL. The ethnic-specific GWAS summary statistics were also combined using the MANTRA (Meta-Analysis of Trans-ethnic Association Studies) package, a meta-analysis software tool allowing for heterogeneity in allelic effects due to differences in LD structure in different ancestry clusters.[24] MANTRA results are reported as log10 Bayes's factors (log_{10}BF). The three sets of ethnic-specific results were furthermore analysed by means of the Han and Eskin RE2 model, a meta-analysis method developed for higher statistical power under heterogeneity.[25] We used the METASOFT 3.0c tool as developed by the Buhm Han laboratories (http://www.buhmhan.com/home) . For the fixed-effects and the RE2 models we applied a genome-wide significant threshold value of P< 5E-08, given the traits under investigation are correlated (**Table S10**). For the MANTRA discovery analyses, a log_{10}BF > 6.1 was considered as a genome-wide significant threshold value.[56]

Replication in human cohorts

The six independent replication studies: the Gutenberg Health Study (GHS cohorts 1 and 2, both EUR ancestry), the Genes and Blood-Clotting Study (GBC, EUR ancestry), the NEO study (EUR ancestry) , the JUPITER trial (EUR ancestry), and the HANDLS study (AFR ancestry) [16–21] (total replication size N= 16,389) provided linear regression results for the novel trait-locus combinations. Their results were meta-analyzed with a fixed effects inverse variance weighted method (METAL) and the RE2 methodology. Additionally, we meta-analyzed replication results with the discovery data using fixed-effects, MANTRA, and RE2 methods. For the replication analyses of the nine individual trait–locus combinations we applied a threshold of P<0.05/9. Additional human replication findings are provided in **Supplemental Data**.

Fine-mapping

We used the MANTRA results to fine-map the regions of trait associated index SNPs. We defined regions by identifying variants within a 1 Mb window around each index SNP (500kb upstream and 500 kb downstream). For each SNP in a region, the posterior probability that this SNP is driving the region's association signal was calculated by dividing the SNP's BF by the summation of the BFs of all SNPs in the region. Credible sets (CS) were subsequently created by sorting the SNPs in each region in descending order based on their BF (so starting with the index SNP since this SNP has the region's largest BF by definition). Going down the sorted list, the SNPs' posterior probabilities were summed until the cumulative value exceeded 99% of the total cumulative posterior probability for all SNPs in the region. The length of a CS was expressed in base pairs. We compared 99% credible sets for the trans-ethnic results and the results of a EUR-only MANTRA analysis.[13,24,26] For the MANTRA fine-mapping analyses, a less stringent threshold value of log10 BF > 5 was applied, because we wanted to include previously identified regions which may not have showed up in the more stringent MANTRA discovery analyses.

Heterogeneity analysis

Heterogeneity of the associations across the different ethnicities was assessed by the I^2 and Cochran's Q statistics as reported by METAL[22] , and the posterior probability of heterogeneity as reported by MANTRA.[24]

Zebrafish gene loss-of-function experiments

Zebrafish *rbpms*, *rbpms2* and *gtf2e2* were identified and confirmed by peptide sequence homology study and gene synteny analysis. For *rbpms*, we relied solely on peptide homology comparison and domain structure since no syntenic region was previously annotated and found by this study. For each morpholino (MO) its design incorporated information about gene structure and translational initiation sites (Gene-Tool Inc., Philomath, OR). MOs targeting each transcript, were injected into single-cell embryos at 1, 3, and 5

ng/embyo doses to find an optimal dose at which there were minimum non-specific toxicity. The step-wise doses also give a range of phenotypes from a hypomorph to a near complete knockdown for most transcripts, which were used to assess the model of an additive model of genetic association. Post-injection, embryos were collected at specified time points, 16-18 ss, 22-26 hpf, and 48 hpf using both standard morphological features of the whole embryo and hours post-fertilization (hpf) to stage to minimize differences in embryonic development staging caused by the MO injection.[28,29] The embryos were then assayed for hematopoietic development by whole-mount in situ hybridization and benzidine staining. We conducted two assays simultaneously for globin transcription and hemoglobin formation. For the globin transcription, developing erythrocytes in the intermediate cell mass of the embryos were assayed by embryonic β-globin 3 expression at the 16 somite stage, or 16-18 hpf.[29] Benzidine staining phenotype was categorized from subtle decrease to complete absence of staining, which was categorized as mild, intermediate or strong effect. Morphologically normal morphants with decreased blood formation were scored for hematopoietic effect.

In zebrafish, the *rbpms* gene was not annotated in the known EST and cDNA databases, although a genomic sequence in the telomeric region on chromosome 7 predicting a coding sequence (80% peptide sequence similarity) was identified. In addition, the synteny between human *RBPMS* and *GTF2E2* genes is not conserved in zebrafish where *rbpms* and *gtf2e2* are located on two separate chromosomes, chromosome 7 and 1, respectively. The *rbpms2* gene was annotated with two paralogs on chromosome 7 (26 Mb away from and centromeric to the true *rbpms* gene) and chromosome 25 of the zebrafish genome. This orthology mapping was confirmed again by this research based on gene synteny and 88 and 91% sequence similarity, respectively for *rbpms2b* and *rbpms2a* to the human *RBPMS2* gene. These two zebrafish *RBPMS2* orthologs have a higher overall sequence similarity to the human *RBPMS* gene than the true zebrafish *rbpms,* but both have a *RBPMS2*-signature stretch of Alanine in the C-terminus of the protein. Therefore, to confirm our *rbpms* orthology study, and to confirm functional conservation of *rbpms* gene in zebrafish, MO individual

knockdown of both *rbpms2a* and *rbpms2b* was also performed in independent experiments, showing much less or no effect by *rbpms2a* knock-down and moderate effect by *rbpms2b* impact on erythropoiesis, suggesting functional compensation of the genes in the *rbpms* family in zebrafish during embryonic erythropoiesis.

Chromatin Immunoprecipitation (ChIP)

For ChIP-seq experiments the following antibodies were used: Gata1 (Santa Cruz sc265X), Gata2 (Santa Cruz sc9008X) and H3K27ac (Abcam ab4729). ChIP experiments were performed as previously described with slight modifications.[57,58] Briefly, 20-30 million cells for each ChIP were crosslinked by the addition of 1/10 volume 11% fresh formaldehyde for 10 min at room temperature. The crosslinking was quenched by the addition of 1/20 volume 2.5M Glycine. Cells were washed twice with ice-cold PBS and the pellet was flash-frozen in liquid nitrogen. Cells were kept at -80°C until the experiments were performed. Cells were lysed in 10 ml of Lysis buffer 1 (50 mM HEPES-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, and protease inhibitors) for 10 min at 4C. After centrifugation, cells were resuspended in 10 ml of Lysis buffer 2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, and protease inhibitors) for 10 min at room temperature. Cells were pelleted and resuspended in 3 ml of Sonication buffer for K562 and U937 and 1 ml for other cells used (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-Deoxycholate, 0.05% Nlauroylsarcosine, and protease Inhibitors) and sonicated in a Bioruptor sonicator for 24-40 cycles of 30s followed by 1min resting intervals. Samples were centrifuged for 10 min at 18,000 g and 1% of TritonX was added to the supernatant. Prior to the immunoprecipitation, 50 ml of protein G beads (Invitrogen 100-04D) for each reaction were washed twice with PBS, 0.5% BSA twice. Finally the beads were resuspended in 250 ml of PBS, 0.5% BSA and 5 mg of each antibody. Beads were rotated for at least 6 hr at 40C and then washed twice with PBS, 0.5% BSA. Cell lysates were added to the beads and incubated at 40C overnight. Beads were washed 1x with (20 mM Tris-HCl (pH 8), 150 mM NaCl, 2mM EDTA, 0.1% SDS, 1%Triton X-100), 1x with (20 mM Tris-HCl

(pH 8), 500 mM NaCl, 2 mM EDTA, 0.1% SDS, 1%Triton X-100), 1x with (10 mM Tris-HCl (pH 8), 250 nM LiCl, 2 mM EDTA, 1% NP40) and 1x with TE and finally resuspended in 200 ml elution buffer (50 mM Tris-Hcl, pH 8.0, 10 mM EDTA and 0.5%–1% SDS) Fifty microliters of cell lysates prior to addition to the beads was kept as input. Crosslinking was reversed by incubating samples at 65°C for at least 6 hr. Afterwards the cells were treated with RNase and proteinase K and the DNA was extracted by Phenol/Chloroform extraction.

ChIP-Seq library Preparation

Briefly, ChIP-Seq libraries were prepared using the following protocol. End repair of immunoprecipitated DNA was performed using the End-It DNA End-Repair kit (Epicentre, ER81050) and incubating the samples at 25°C for 45 min. End repaired DNA was purified using AMPure XP Beads (1.8X of the reaction volume) (Agencourt AMPure XP – PCR purification Beads, BeckmanCoulter, A63881) and separating beads using DynaMag-96 Side Skirted Magnet (Life Technologies, 12027). A tail was added to the end-repaired DNA using NEB Klenow Fragment Enzyme (3'-5' exo, M0212L), 1X NEB buffer 2 and 0.2 mM dATP (Invitrogen, 18252-015) and incubating the reaction mix at 37°C for 30 min. A-tailed DNA was cleaned up using AMPure beads (1.8X of reaction volume). Subsequently, cleaned up dA-tailed DNA went through Adaptor ligation reaction using Quick Ligation Kit (NEB, M2200L) following manufacturer's protocol. Adaptor-ligated DNA was first cleaned up using AMPure beads (1.8X of reaction volume), eluted in 100µl and then size-selected using AMPure beads (0.9X of the final supernatant volume, 90 µl). Adaptor ligated DNA fragments of proper size were enriched with PCR reaction using Fusion High-Fidelity PCR Master Mix kit (NEB, M0531S) and specific index primers supplied in NEBNext Multiplex Oligo Kit for Illumina (Index Primer Set 1, NEB, E7335L). Conditions for PCR used are as follows: 98 °C , 30 sec; [98°C, 10 sec; 65 °C, 30 sec; 72 °C, 30 sec] X 15 to 18 cycles; 72°C, 5 min; hold at 4 °C. PCR enriched fragments were further size selected by running the PCR reaction mix in 2% low-molecular weight agarose gel (Bio-Rad, 161- 3107) and subsequently purifying them using QIAquick Gel Extraction Kit (28704). Libraries were eluted in 25µl elution buffer. After

measuring concentration in Qubit, all the libraries went through quality control analysis using an Agilent Bioanalyzer. Samples with proper size (250-300 bp) were selected for next generation sequencing using Illumina Hiseq 2000 or 2500 platform.

ChIP-Seq data analysis

Alignment and Visualization ChIP-Seq reads were aligned to the human reference genome (hg19) using bowtie with parameters -k 2 -m 2 -S.[59] WIG files for display were created using MACS[60] with parameters -w -S --space=50 --nomodel --shiftsize=200 and were displayed in IGV.[61,62]

Peak and Bound Gene Identification

High-confidence peaks of ChIP-Seq signal were identified using MACS with parameters --keepdup= auto -p 1e-9 and corresponding input control. Bound genes are RefSeq genes that contact a MACS-defined peak between -10000bp from the TSS and +5000bp from the TES.

Assay for Transposase Accessible Chromatin (ATACseq)

CD34+ cells were expanded and differentiated using the protocol mentioned above. Before collection, cells were treated with 25 ng/ml hrBMP4 for 2 hr. 5 x 10^4 cells per differentiation stage were harvested by spinning at 500 x g for 5 min, 4° C. Cells were washed once with 50 µL of cold 1X PBS and spinned down at 500 x g for 5 min, 4° C. After discarding supernatant, cells were lysed using 50 µL cold lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-360) and spun down immediately at 500 x g for 10 mins, 4 ºC. Then the cells were precipitated and kept on ice and subsequently resuspended in 25 µL 2X TD Buffer (Illumina Nextera kit), 2.5 µL Transposase enzyme (Illumina Nextera kit, 15028252) and 22.5 µL Nuclease-free water in a total of 50 µL reaction for 1 hr at 37° C. DNA was then purified using Qiagen MinElute PCR purification kit (28004) in a final volume of 10 µL. Libraries were constructed according to Illumina protocol using the DNA treated

with transposase, NEB PCR master mix, Sybr green, universal and library-specific Nextera index primers. The first round of PCR was performed under the following conditions: 72° C, 5 min; 98° C, 30 sec; [98°C, 10 sec; 63 °C, 30 sec; 72 °C, 1 min] X 5 cycles; hold at 4°C. Reactions were kept on ice and using a 5 µL reaction aliquot, the appropriate number of additional cycles required for further amplification was determined in a side qPCR reaction: 98 °C , 30 sec; [98 °C, 10 sec; 63 °C, 30 sec; 72 °C, 1 min] X 20 cycles; hold at 4 °C. Upon determining the additional number of PCR cycles required further for each sample, library amplification was conducted using the following conditions: 98°C, 30 sec; [98°C,10 sec; 63°C, 30 sec; 72 °C, 1 min] X appropriate number of cycles; hold at 4°C. Libraries prepared went through quality control analysis using an Agilent Bioanalyzer. Samples with appropriate nucleosomal laddering profiles were selected for next generation sequencing using Illumina Hiseq 2500 platform.

ATACseq data analysis

All human ChIP-Seq datasets were aligned to build version NCBI37/HG19 of the human genome using Bowtie2 (version 2.2.1)[59] with the following parameters: --end-to-end, -N0, -L20. We used the MACS2 version 2.1.0 [60] peak finding algorithm to identify regions of ATAC-Seq peaks, with the following parameter --nomodel --shift -100 --extsize 200. A q-value threshold of enrichment of 0.05 was used for all datasets.

Mouse QTL mapping

The methods for the original mouse cross have been previously described.[30] Briefly, mice from 12 different strains were inter-crossed[30] and the same erythrocyte traits we have studied by GWAS were measured in peripheral blood. The Jackson Laboratory Animal Care and Use Committee approved all protocols. The number of markers genotyped per cross varied by the platform used, and the total number per cross is provided in **Table S9**. QTL analysis was performed for each erythrocyte trait using R/qtl v1.07-12 http://www.rqtl.org).[63] Genetic map positions of all markers used were updated to the new mouse genetic map

using online mouse map converter tool at http://cgd.jax.org.[64] All phenotypic data were ranked-Z transformed to approximate the normal distribution prior to analysis. The QTL analysis was performed as a genome-wide scan with sex as an additive covariate. Permutation testing (1000 permutations) was used to determine significance, and LOD scores greater than the 95th percentile ($P < 0.05$) were considered significant. QTL confidence intervals were determined by the posterior probability.[65,66] For each candidate region in the mouse, the coordinates were obtained from the Mouse Genome Database, which is part of Mouse Genome Informatics (MGI), using the 'Genes and Markers' query (http://www.informatics.jax.org/marker/). Protein coding genes, non-coding RNA genes, and unclassified genes were queried.

Pleiotropy analysis

To further relate the zebrafish findings in the *rbpms* gene family back to the human association data and to explore possible pleiotropy which might suggest effects on an earlier hematopoietic progenitor cell, we evaluated *RBPMS* association with WBC traits and *RBPMS2* association with RBC and WBC traits(**Supplemental Data)**.

Supplemental Data

Supplemental data include individual cohort descriptions, methods replication findings, 69 tables, and 123 figures

Acknowledgements

Funding for Age, Gene/ Environment Susceptibility Reykjavik Study (AGES)was made possible by NIA/NIH contract AG000932-2 (2009) Characterization of Normal Genomic Variability. The Age, Gene/ Environment Susceptibility Reykjavik Study is funded by NIH contract N01-AG-12100, the NIA Intramural Research Program, Hjartavernd (the Icelandic Heart Association) and the Althingi (the Icelandic Parliament).

Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

Cardiovascular Health Study: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants U01HL080295, R01HL087652, R01HL105756, R01HL103612, and R01HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org.

The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of

Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The National Heart, Lung, and Blood Institute's Framingham Heart Study is a joint project of the National Institutes of Health and Boston University School of Medicine and was supported by the National Heart, Lung, and Blood Institute's Framingham Heart Study (contract No. N01-HC-25195) and its contract with Affymetrix for genotyping services (contract No. N02-HL-6-4278). Analyses reflect the efforts and resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. A portion of this research was conducted using the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at the Boston University School of Medicine and Boston Medical Center. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

The Health ABC Study was supported in part by the Intramural Research Program of the NIH, National Institute on Aging, NIA contracts N01AG62101, N01AG62103 and N01AG 62106. The GWAS was funded by NIA grant 1R01AG032098- 01A1 to Wake Forest University Health Sciences and genotyp- ing services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal con- tract from the National Institutes of Health to The Johns Hopkins University (contract number HHSN268200782096C).

The PROSPER study was supported by an investigator initiated grant obtained from Bristol-Myers Squibb. Prof. Dr. J. W. Jukema is an Established Clinical Investigator of the

Netherlands Heart Foundation (grant 2001 D 032). Support for genotyping was provided by the seventh framework program of the European commission (grant 223004) and by the Netherlands Genomics Initiative (Netherlands Consortium for Healthy Aging grant 050-060-810).

The InChianti Study was supported as a 'targeted project' (ICS 110.1RS97.71) by the Italian Ministry of Health, by the US Na- tional Institute on Aging (contracts N01-AG-916413, N01- AG-821336, 263 MD 9164 13 and 263 MD 821336) and in part by the Intramural Research Program, National Institute on Aging, National Institutes of Health, USA.

The generation and management of GWAS genotype data for the Rotterdam Study (RS I, RS II, RS III) was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. The GWAS datasets are supported by the Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project nr. 050-060-810. We thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera and Marjolein Peters, MSc, and Carolina Medina-Gomez, MSc, for their help in creating the GWAS database, and Karol Estrada, PhD, Yurii Aulchenko, PhD, and Carolina Medina-Gomez, MSc, for the creation and analysis of imputed data. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

Funding for COGENT was obtained through the Broad Institute (N01-HC- 65226) to create this genotype/phenotype database for wide dissemination to the biomedical research community.

Coronary Artery Risk in Young Adults (CARDIA): University of Alabama at Birmingham (N01-HC- 48047), University of Minnesota (N01-HC-48048), North- western University (N01-HC-48049), Kaiser Foundation Research Institute (N01-HC-48050), University of Alabama at Birmingham (N01-HC-95095), Tufts-New England Medical Center (N01-HC-45204), Wake Forest University (N01-HC- 45205), Harbor-UCLA Research and Education Institute (N01- HC-05187), University of California, Irvine (N01-HC-45134 and N01-HC-95100).

Jackson Heart Study (JHS): Contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities. .

Healthy Aging in Neighborhoods of Diversity across the Life Span Study (HANDLS): This research was supported by the Intramural Research Program of the NIH, National Institute on Aging and the National Center on Minority Health and Health Disparities (intra-mural project # Z01-AG000513 and human subjects protocol # 2009-149).

Health ABC: This research was supported by NIA contracts N01AG62101, N01AG62103 and N01AG62106. The GWAS was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University (contract number HHSN268200782096C). This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

GeneSTAR: This research was supported by the National Heart, Lung, and Blood Institute (NHLBI) through the PROGENI (U01 HL72518) and STAMPEED (R01 HL087698-01) consortia. Additional support was provided by grants from the NIH/National In- stitute of Nursing Research (R01 NR08153) and the NIH/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center.

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, US Department of Health and Human Services, through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32 and 44221.

Funding for RIKEN and the BioBank Japan Project was supported by Ministry of Education, Culture, Sports, Science and Technology, Japan.

The Gutenberg Health Study is funded through the government of Rhineland-Palatinate („Stiftung Rheinland-Pfalz für Innovation“, contract AZ 961-386261/733), the research programs “Wissen schafft Zukunft” and “Center for Translational Vascular Biology (CTVB)” of the Johannes Gutenberg-University of Mainz, and its contract with Boehringer Ingelheim and PHILIPS Medical Systems, including an unrestricted grant for the Gutenberg Health Study. VG PSW are funded by the Federal Ministry of Education and Research (BMBF 01EO1503). TZ and PSW are PIs of the German Center for Cardiovascular Research. The remaining authors have nothing to declare.

The Genes and Blood Clotting Study was supported by the National Institute of Health grants R37HL039693 (K.C.D., D.G.) and RO1HL112642 (A.B.O., K.C.D.,J.Z.L., D.G.). Additionally, David Ginsburg is a Howard Hughes Medical Institute investigator.

The authors of the NEO study thank all individuals who participated in the Netherlands Epidemiology in Obesity study, all participating general practitioners for inviting eligible

participants and all research nurses for collection of the data. We thank the NEO study group, Pat van Beelen, Petra Noordijk and Ingeborg de Jonge for the coordination, lab and data management of the NEO study. The genotyping in the NEO study was supported by the Centre National de Génotypage (Paris, France), headed by Jean-Francois Deleuze. The NEO study is supported by the participating Departments, the Division and the Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. Dennis Mook-Kanamori is supported by Dutch Science Organization (ZonMW-VENI Grant 916.14.023).

The JUPITER trial and the genotyping were supported by AstraZeneca.

The development of the software package MANTRA was performed by APM, a Wellcome Trust Senior Research Fellow in Basic Biomedical Science (grant numbers WT098017, WT090532 and WT064890).

Professor Luanne L. Peters (LLP) is supported by NIH grants HL085480 and DK100692. S.K.Ganesh is supported by the Doris Duke Charitable Foundation and NIH HL122684 grants. Yukinori Okada was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI grant numbers 15H05911, 15H05670, 15K14429, the Japan Science and Technology Agency (JST), Mochida Memorial Foundation for Medical and Pharmaceutical Research, Takeda Science Foundation, Gout Research Foundation, the Tokyo Biochemical Research Foundation, and the Japan Rheumatism Foundation. Robert J. Klein was supported by grant number U01 HG007033 from NHGRI.

The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Disclosure statement

Bruce M. Psaty serves on the DSMB of a clinical trial funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale Open Data Access project funded by Johnson & Johnson.

The other authors declare no relevant financial interests.

References

1. Koury, M.J. (2014). Abnormal erythropoiesis and the pathophysiology of chronic anemia. Blood Rev. *28*, 49–66.

2. Whitfield, J.B., Martin, N.G., and Rao, D.C. (1985). Genetic and environmental influences on the size and number of cells in the blood. Genet. Epidemiol. *2*, 133–144.

3. Evans, D.M., Frazer, I.H., and Martin, N.G. (1999). Genetic and environmental causes of variation in basal levels of blood cells. Twin Res. Hum. Genet. *2*, 250–257.

4. Lin, J.-P., O'Donnell, C.J., Jin, L., Fox, C., Yang, Q., and Cupples, L.A. (2007). Evidence for linkage of red blood cell size and count: Genome-wide scans in the Framingham Heart Study. Am. J. Hematol. *82*, 605–610.

5. Guindo, A., Fairhurst, R.M., Doumbo, O.K., Wellems, T.E., and Diallo, D.A. (2007). X-Linked G6PD Deficiency Protects Hemizygous Males but Not Heterozygous Females against Severe Malaria. PLoS Med. *4*, e66.

6. Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., et al. (2001). Haplotype Diversity and Linkage Disequilibrium at Human G6PD: Recent Origin of Alleles That Confer Malarial Resistance. Science *293*, 455–462.

7. Lo, K.S., Wilson, J.G., Lange, L.A., Folsom, A.R., Galarneau, G., Ganesh, S.K., Grant, S.F.A., Keating, B.J., McCarroll, S.A., Mohler, E.R., et al. (2011). Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. Hum. Genet. *129*, 307–317.

8. Ganesh, S.K., Zakai, N.A., van Rooij, F.J.A., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.-H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat. Genet. *41*, 1191–1198.

9. Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. *41*, 1182–1190.

10. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.

11. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat. Genet. *42*, 210–215.

12. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum. Mol. Genet. *22*, 2529–2538.

13. Franceschini, N., van Rooij, F.J.A., Prins, B.P., Feitosa, M.F., Karakas, M., Eckfeldt, J.H., Folsom, A.R., Kopp, J., Vaez, A., Andrews, J.S., et al. (2012). Discovery and Fine Mapping of Serum Protein Loci through Transethnic Meta-analysis. Am. J. Hum. Genet. *91*, 744–753.

14. Nalls, M.A., Couper, D.J., Tanaka, T., van Rooij, F.J.A., Chen, M.-H., Smith, A.V., Toniolo, D., Zakai, N.A., Yang, Q., Greinacher, A., et al. (2011). Multiple Loci Are Associated with White Blood Cell Phenotypes. PLoS Genet *7*, e1002113.

15. Chen, P., Takeuchi, F., Lee, J.-Y., Li, H., Wu, J.-Y., Liang, J., Long, J., Tabara, Y., Goodarzi, M.O., Pereira, M.A., et al. (2014). Multiple Nonglycemic Genomic Loci Are Newly Associated With Blood Level of Glycated Hemoglobin in East Asians. Diabetes *63*, 2551–2562.

16. Wild, D.P.S., Zeller, T., Beutel, M., Blettner, M., Dugi, K.A., Lackner, K.J., Pfeiffer, N., Münzel, T., and Blankenberg, P.D.S. (2012). Die Gutenberg Gesundheitsstudie. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz *55*, 824–830.

17. Desch, K.C., Ozel, A.B., Siemieniak, D., Kalish, Y., Shavit, J.A., Thornburg, C.D., Sharathkumar, A.A., McHugh, C.P., Laurie, C.C., Crenshaw, A., et al. (2013). Linkage analysis identifies a locus for plasma von Willebrand factor undetected by genome-wide association. Proc. Natl. Acad. Sci. U. S. A. *110*, 588–593.

18. Mutsert, R. de, Heijer, M. den, Rabelink, T.J., Smit, J.W.A., Romijn, J.A., Jukema, J.W., Roos, A. de, Cobbaert, C.M., Kloppenburg, M., Cessie, S. le, et al. (2013). The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. Eur. J. Epidemiol. *28*, 513–523.

19. Ridker, P.M. (2003). Rosuvastatin in the Primary Prevention of Cardiovascular Disease Among Patients With Low Levels of Low-Density Lipoprotein Cholesterol and Elevated High-Sensitivity C-Reactive Protein Rationale and Design of the JUPITER Trial. Circulation *108*, 2292–2297.

20. Qayyum, R., Snively, B.M., Ziv, E., Nalls, M.A., Liu, Y., Tang, W., Yanek, L.R., Lange, L., Evans, M.K., Ganesh, S., et al. (2012). A Meta-Analysis and Genome-Wide Association Study of Platelet Count and Mean Platelet Volume in African Americans. PLoS Genet *8*, e1002491.

21. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., et al. (2011). Genome-Wide Association Study of White Blood Cell Count in 16,388 African Americans: the Continental Origins and Genetic Epidemiology Network (COGENT). PLoS Genet *7*, e1002108.

22. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.

23. Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic Control, a New Approach to Genetic-Based Association Studies. Theor. Popul. Biol. *60*, 155–166.

24. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.

25. Han, B., and Eskin, E. (2011). Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. Am. J. Hum. Genet. *88*, 586–598.

26. The Wellcome Trust Case Control Consortium, Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet. *44*, 1294–1301.

27. The ENCODE Project Consortium (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. *9*,.

28. Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., and Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Dev. Dyn. *203*, 253–310.

29. Huang, H.-T., Kathrein, K.L., Barton, A., Gitlin, Z., Huang, Y.-H., Ward, T.P., Hofmann, O., Dibiase, A., Song, A., Tyekucheva, S., et al. (2013). A network of epigenetic regulators guides developmental haematopoiesis in vivo. Nat. Cell Biol. *15*, 1516–1525.

30. Peters, L.L., Shavit, J.A., Lambert, A.J., Tsaih, S.-W., Li, Q., Su, Z., Leduc, M.S., Paigen, B., Churchill, G.A., Ginsburg, D., et al. (2010). Sequence variation at multiple loci influences red cell hemoglobin concentration. Blood *116*, e139–e149.

31. Chambers, J.C., Zhang, W., Li, Y., Sehmi, J., Wass, M.N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M.I., Peltonen, L., et al. (2009). Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nat. Genet. *41*, 1170–1172.

32. Ding, K., de Andrade, M., Manolio, T.A., Crawford, D.C., Rasmussen-Torvik, L.J., Ritchie, M.D., Denny, J.C., Masys, D.R., Jouni, H., Pachecho, J.A., et al. (2013). Genetic Variants That Confer Resistance to Malaria Are Associated with Red Blood Cell Traits in African-Americans: An Electronic Medical Record-based Genome-Wide Association Study. G3 GenesGenomesGenetics *3*, 1061–1068.

33. Kullo, I.J., Ding, K., Jouni, H., Smith, C.Y., and Chute, C.G. (2010). A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. PLoS ONE *5*, e13011.

34. Li, J., Glessner, J.T., Zhang, H., Hou, C., Wei, Z., Bradfield, J.P., Mentch, F.D., Guo, Y., Kim, C., Xia, Q., et al. (2013). GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Hum. Mol. Genet. *22*, 1457–1464.

35. Pistis, G., Okonkwo, S.U., Traglia, M., Sala, C., Shin, S.-Y., Masciullo, C., Buetti, I., Massacane, R., Mangino, M., Thein, S.-L., et al. (2013). Genome Wide Association Analysis of a Founder Population Identified TAF3 as a Gene for MCHC in Humans. PLoS ONE *8*,.

36. the CHARGE Consortium Hematology Working Group (2016). Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. Nat. Genet. *48*, 867–876.

37. Richardson, K., Louie-Gao, Q., Arnett, D.K., Parnell, L.D., Lai, C.-Q., Davalos, A., Fox, C.S., Demissie, S., Cupples, L.A., Fernandez-Hernando, C., et al. (2011). The PLIN4 Variant rs8887 Modulates Obesity Related Phenotypes in Humans through Creation of a Novel miR-522 Seed Site. PLOS ONE *6*, e17944.

38. Williams, Z., Ben-Dov, I.Z., Elias, R., Mihailovic, A., Brown, M., Rosenwaks, Z., and Tuschl, T. (2013). Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. Proc. Natl. Acad. Sci. U. S. A. *110*, 4255–4260.

39. Shimamoto, A., Kitao, S., Ichikawa, K., Suzuki, N., Yamabe, Y., Imamura, O., Tokutake, Y., Satoh, M., Matsumoto, T., Kuromitsu, J., et al. (1996). A unique human gene that spans over 230 kb in the human chromosome 8p11-12 and codes multiple family proteins sharing RNA-binding motifs. Proc. Natl. Acad. Sci. U. S. A. *93*, 10913–10917.

40. Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. Wiley Interdiscip. Rev. RNA *3*, 159–177.

41. Trakarnsanga, K., Wilson, M.C., Griffiths, R.E., Toye, A.M., Carpenter, L., Heesom, K.J., Parsons, S.F., Anstee, D.J., and Frayne, J. (2014). Qualitative and Quantitative Comparison of the Proteome of Erythroid Cells Differentiated from Human iPSCs and Adult Erythroid Cells by Multiplex TMT Labelling and NanoLC-MS/MS. PLoS ONE *9*, e100874.

42. Kingsley, P.D., Greenfest-Allen, E., Frame, J.M., Bushnell, T.P., Malik, J., McGrath, K.E., Stoeckert, C.J., and Palis, J. (2013). Ontogeny of erythroid gene expression. Blood *121*, e5–e13.

43. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C., and Melton, D.A. (2002). "Stemness": transcriptional profiling of embryonic and adult stem cells. Science *298*, 597–600.

44. Georgantas, R.W., Tanadve, V., Malehorn, M., Heimfeld, S., Chen, C., Carr, L., Martinez-Murillo, F., Riggins, G., Kowalski, J., and Civin, C.I. (2004). Microarray and Serial Analysis of Gene Expression Analyses Identify Known and Novel Transcripts Overexpressed in Hematopoietic Stem Cells. Cancer Res. *64*, 4434–4441.

45. Wagner, W., Ansorge, A., Wirkner, U., Eckstein, V., Schwager, C., Blake, J., Miesala, K., Selig, J., Saffrich, R., Ansorge, W., et al. (2004). Molecular evidence for stem cell function of the slow-dividing fraction among human hematopoietic progenitor cells by genome-wide analysis. Blood *104*, 675–686.

46. Sun, Y., Ding, L., Zhang, H., Han, J., Yang, X., Yan, J., Zhu, Y., Li, J., Song, H., and Ye, Q. (2006). Potentiation of Smad-mediated transcriptional activation by the RNA-binding protein RBPMS. Nucleic Acids Res. *34*, 6314–6326.

47. He, W., Dorn, D.C., Erdjument-Bromage, H., Tempst, P., Moore, M.A.S., and Massagué, J. (2006). Hematopoiesis Controlled by Distinct TIF1γ and Smad4 Branches of the TGFβ Pathway. Cell *125*, 929–941.

48. Shin, S.-Y., Fauman, E.B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., et al. (2014). An atlas of genetic influences on human blood metabolites. Nat. Genet. *46*, 543–550.

49. Comuzzie, A.G., Cole, S.A., Laston, S.L., Voruganti, V.S., Haack, K., Gibbs, R.A., and Butte, N.F. (2012). Novel Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic Population. PLoS ONE *7*, e51954.

50. International Parkinson Disease Genomics Consortium, Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.-M., Saad, M., Simón-Sánchez, J., Schulte, C., Lesage, S., et al. (2011). Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet *377*, 641–649.

51. Otsuki, M., Fukami, K., Kohno, T., Yokota, J., and Takenawa, T. (1999). Identification and Characterization of a New Phospholipase C-like Protein, PLC-L2. Biochem. Biophys. Res. Commun. *266*, 97–103.

52. Sankaran, V.G., Ludwig, L.S., Sicinska, E., Xu, J., Bauer, D.E., Eng, J.C., Patterson, H.C., Metcalf, R.A., Natkunam, Y., Orkin, S.H., et al. (2012). Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. Genes Dev. *26*, 2075–2087.

53. Keller, M.F., Reiner, A.P., Okada, Y., Rooij, F.J.A. van, Johnson, A.D., Chen, M.-H., Smith, A.V., Morris, A.P., Tanaka, T., Ferrucci, L., et al. (2014). Trans-ethnic meta-analysis of white blood cell phenotypes. Hum. Mol. Genet. ddu401.

54. Dastani, Z., Hivert, M.-F., Timpson, N., Perry, J.R.B., Yuan, X., Scott, R.A., Henneman, P., Heid, I.M., Kizer, J.R., Lyytikainen, L.-P., et al. (2012). Novel Loci for Adiponectin Levels and Their Influence on Type 2 Diabetes and Metabolic Traits: A Multi-Ethnic Meta-Analysis of 45,891 Individuals. PLoS Genet. *8*,.

55. Liu, C.-T., Buchkovich, M.L., Winkler, T.W., Heid, I.M., Borecki, I.B., Fox, C.S., Mohlke, K.L., North, K.E., and Cupples, L.A. (2014). Multi-ethnic fine-mapping of 14 central adiposity loci. Hum. Mol. Genet. *23*, 4738–4744.

56. Wang, X., Chua, H.-X., Chen, P., Ong, R.T.-H., Sim, X., Zhang, W., Takeuchi, F., Liu, X., Khor, C.-C., Tay, W.-T., et al. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. Hum. Mol. Genet. ddt064.

57. Lee, T.I., Johnstone, S.E., and Young, R.A. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat. Protoc. *1*, 729–748.

58. Trompouki, E., Bowman, T.V., Lawton, L.N., Fan, Z.P., Wu, D.-C., DiBiase, A., Martin, C.S., Cech, J.N., Sessa, A.K., Leblanc, J.L., et al. (2011). Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. Cell *147*, 577–589.

59. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

60. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

61. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative Genomics Viewer. Nat. Biotechnol. *29*, 24–26.

62. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

63. Broman, K.W., Wu, H., Sen, Ś., and Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics *19*, 889–890.

64. Cox, A., Ackert-Bicknell, C.L., Dumont, B.L., Ding, Y., Bell, J.T., Brockmann, G.A., Wergedal, J.E., Bult, C., Paigen, B., Flint, J., et al. (2009). A new standard genetic map for the laboratory mouse. Genetics *182*, 1335–1344.

65. Churchill, G.A., and Doerge, R.W. (1994). Empirical Threshold Values for Quantitative Trait Mapping. Genetics *138*, 963–971.

66. Sen, Ś., and Churchill, G.A. (2001). A Statistical Framework for Quantitative Trait Mapping. Genetics *159*, 371–387.

Figure Legends

Figure 1: Fine mapping of the chromosome 8 *RBPMS*/*GTF2E2* locus. 99% credible sets (red dots) around the top hit rs2979489 (red diamond). European Ancestry MANTRA analyses (upper panels) for MCH (left) and MCV (right) are shown, compared to 99% credible sets of the multi-ethnic MANTRA analyses (bottom panels, MCH on the left and MCV on the right).

Figure 2. rs2979489 is localized to a potential regulatory site that involves in transition binding of GATA 2 to GATA1 during erythrocyte differentiation. Top panel. Gene-track view of rs2979489 location in the *RBPMS*/*GTF2E2* gene region. Bottom left Panel: Gene-track of RBPMS gene showing overlap of GATA2, GATA1 and ATACseq peaks (red, blue and green, respectively) during human erythroid differentiation. Bottom right Panel: Overlap of ATACseq (green) and H3K27ac ChIPseq (black) during differentiation at the region proximal to the SNP rs2979489. The grey horizontal line indicates the position of SNP rs 2979489. D0 - Day 0, H6 – Hour 6, D3 – Day3, D4 – Day 4 and D5 – Day 5 of erythroid differentiation time-course post-induction of differentiation.

Figure 3: **Loss-of-function analysis of the *RBPMS, RBPMS2,* and GTF2E2 orthologues in zebrafish**. After injection of 0-3 ng ATG- and splicing-morpholinos against the *RBPMS* zebrafish orthologue, o-dianisidine/benzidine staining in embryos at 48 hours post fertilization (hpf) (A-E right panels) and embryonic βe3 globin expression (A-E left panels) in embryos at 16-18 somite stage (ss) are significantly decreased, indicating a dose-dependent disruption in erythropoiesis in the experimentally treated embryos as compared to controls. Representative results are shown for the *RBPMS* orthologue in embryos (E panel) as well as for *rbpms2a* (C panel) and *rbpms2b* (D panel) at higher doses. Injections of morpholino against the zebrafish *GTF2E2* orthologue (B panel) also at a higher dose show no significant effect on βe3 globin expression at 16-18 ss and o-dianisidine/benzidine staining at 48 hpf. Expression of vascular marker gene, *kdrl* (A-E middle

panels), is relatively normal in all morpholino-injected embryos at 24-26 hpf, suggesting grossly normal development of cells in other organs..

Table 1: Novel findings from the METAL and MANTRA trans-ethnic analyses.

						METAL		MANTRA		RE2
Trait	SNP	Chr	Gene	c/nc	N	Effect (SE)	P	$Log_{10}BF$	posthg	P
Hb	rs2299433	7	*MET*	T/C	63091	0.041 (0.008)	6.16E-08	6.195	0.027	1.20E-07
Hct	rs6430549	2	*TMEM163* / ACMSD	A/G	71647	0.103 (0.018)	4.96E-09	7.408	0.120	8.46E-09
Hct	rs2299433	7	*MET*	T/C	63532	0.102 (0.019)	5.66E-08	6.199	0.099	9.87E-08
MCH	rs2060597	3	*PLCL2*	T/C	38836	0.006 (0.001)	4.18E-10	8.178	0.009	9.75E-10
MCH	rs2979489	8	*RBPMS*	A/G	37531	-0.002 (0.001)	8.89E-05	9.723	1.000	1.19E-12
MCV	rs10929547	2	*ID2*	A/C	50870	-0.002 (0.0003)	2.50E-09	7.977	0.007	2.14E-09
MCV	rs9821630	3	*PLCL2*	A/G	48697	-0.002 (0.0004)	6.86E-09	7.864	0.004	2.44E-09
MCV	rs2979489	8	*RBPMS*	A/G	48697	-0.002 (0.0004)	7.24E-09	7.961	0.003	1.65E-09
MCV	rs6121246	20	*FOXS1*	T/C	49896	0.003 (0.001)	4.05E-07	6.296	0.003	8.31E-08

Chr = chromosome number ; c/nc = coding/non-coding allele ; N = number of participants ; SE = standard error ; P = p-value ; $Log_{10}BF$ = Logaritm of Bayes Factor ; posthg = posterior probability of heterogeneity

Table 2: Fine mapping of a novel locus identified in European ancestry meta-analysis by MANTRA trans-ethnic analysis.

			EUR				Multi-ethnic			
Trait	Chr	Gene	topSNP	log_{10}BF	n_SNPs	width (bp)	topSNP	log_{10}BF	n_SNPs	width (bp)
MCH	8	*RBPMS*	rs2979502	6.32982	21	241480	rs2979489	9.72267	1	1
MCV	8	*RBPMS*	rs2979489	6.13733	11	241480	rs2979489	7.96132	1	1

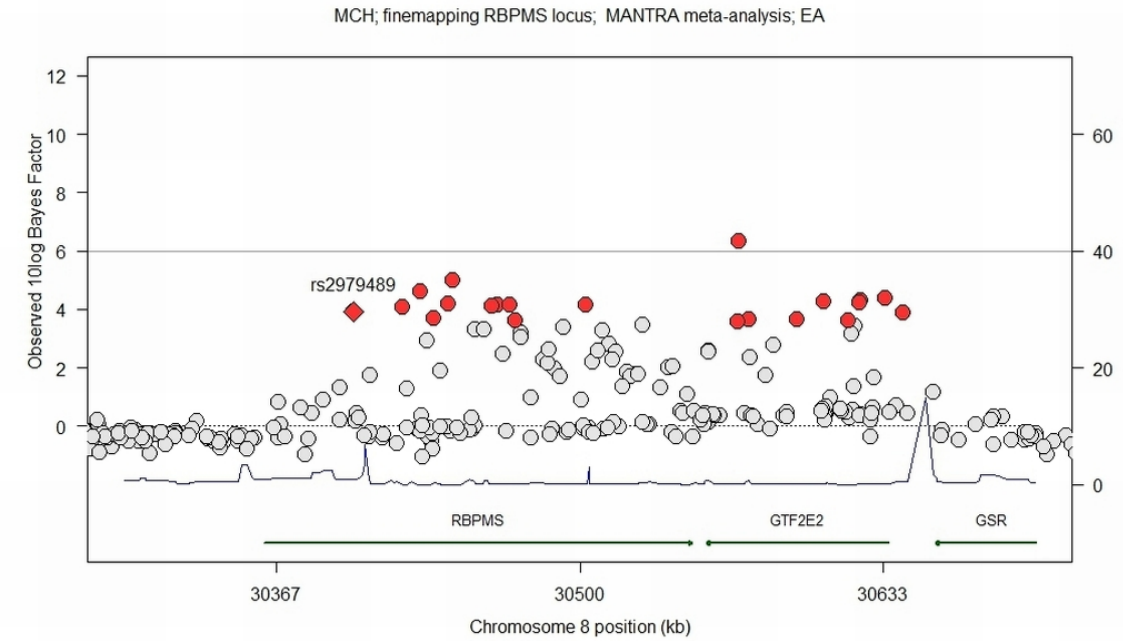
Chr = chromosome number ; Log_{10}BF = Logaritm of Bayes Factor ; n_SNPs : number of SNPs in the region

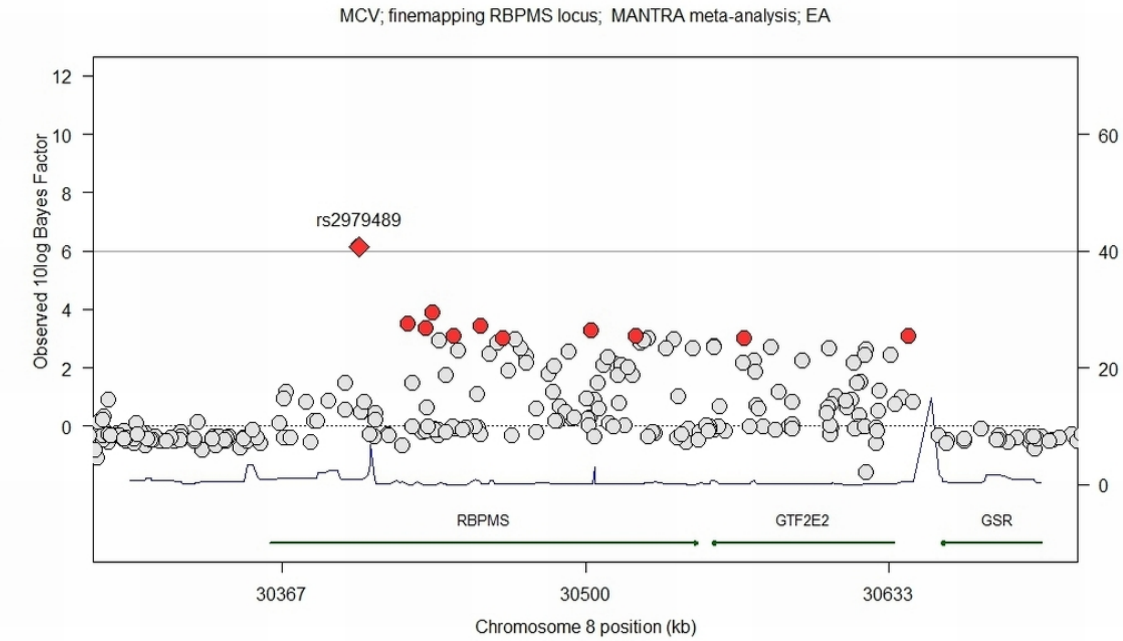
Table 3 Mouse QTL validation of novel MANTRA trans-ethnic findings.

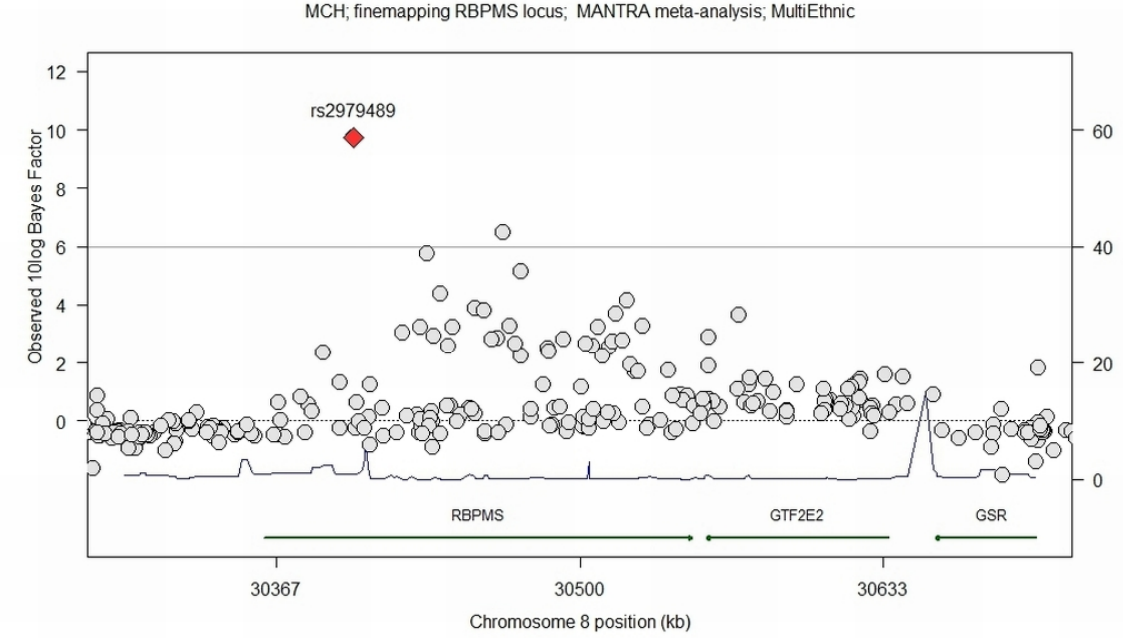
Trait	Chr	Gene	Human (hg18 / Build 36) (chromosome:position)	Mouse (37 mm9) (chromosome:position)		Significant (bold) and Suggestive Mouse QTL [b] peak (95% CI) (Mb)	LOD
Hct	2	*TMEM163*/*ACMSD*	chr2:135196450-135438613	chr1:129581372-129711586	[a]	**141.0 (54.8 – 158.9)**	**3.72**
Hct	4	*SHROOM3*	chr4:77586311-77629342	chr5:93112461-93394344	[b]	46.0 (19.6-106.5)	2.34
Hct	7	*MET*	chr7:116118114-116131947	chr6:17432318-17447418	[a]	37.6 (6.6 – 127.9)	2.75
MCH	8	*RBPMS*	chr8:30400375-30400375	chr8:34893115-35040335	[b]	**78.9 (28.0-96.1)**	**3.98**
MCV	3	*PLCL2*	chr3:16860239-16945942	chr17:50604848-50698773	[a]	**46.0 (28.6 – 55.3)**	**5.46**
MCV	20	*FOXS1*	chr20:29684484-29897013	chr2:152576419-152758874	[a]	**170.1 (147.6 – 179.3)**	**4.69**

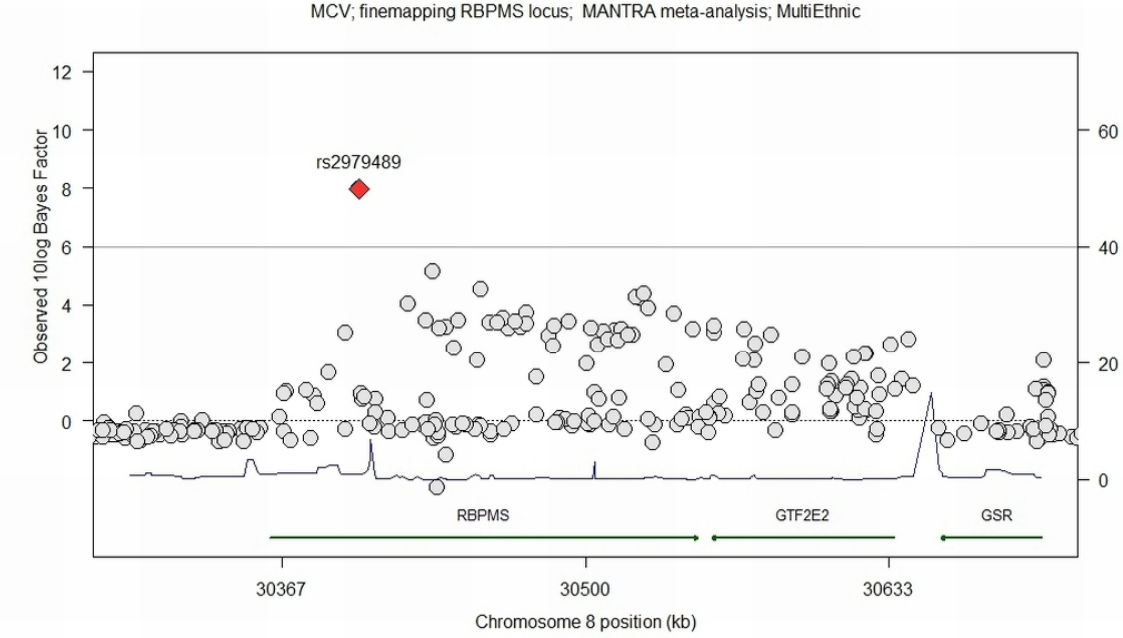
[a] within the corresponding human interval (+/- 250 kb)

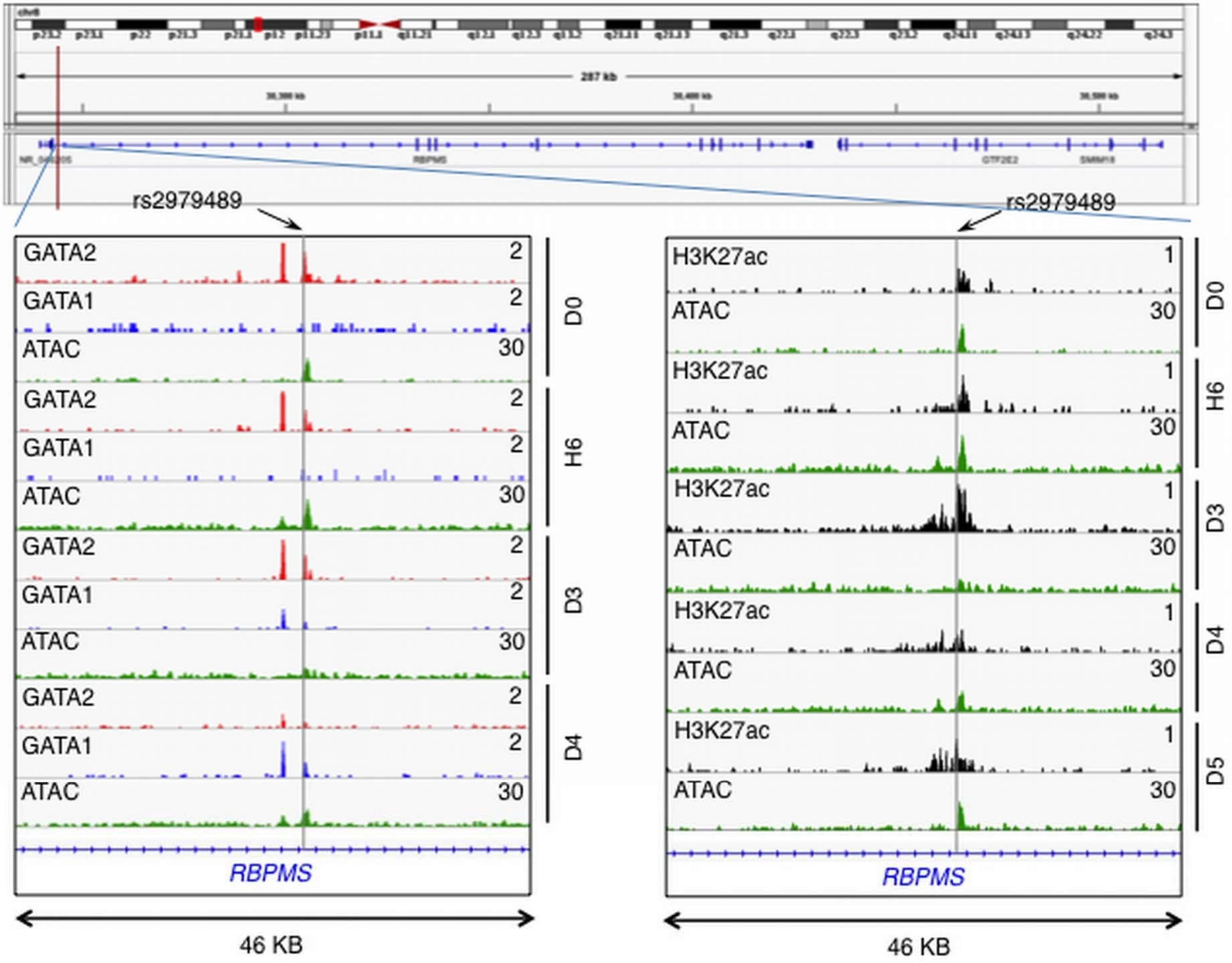
[b] gene found in a significant or suggestive 95% CI mouse QTL, not corresponding to the human interval

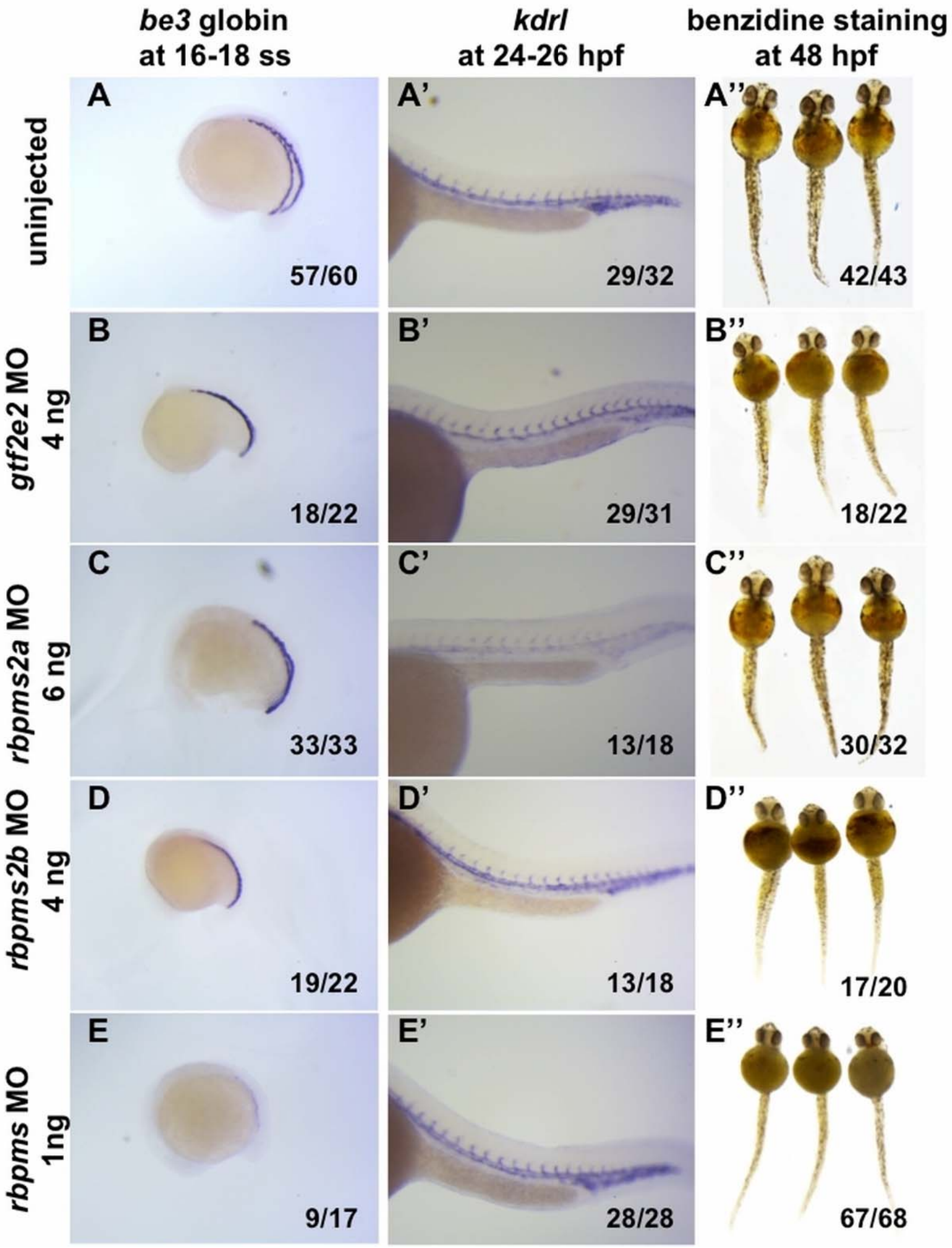

MCH; finemapping RBPMS locus; MANTRA meta-analysis; EA
Observed 10log Bayes Factor
rs2979489
RBPMS
GTF2E2
GSR
30367
30500
30633
Chromosome 8 position (kb)


MCV; finemapping RBPMS locus; MANTRA meta-analysis; EA
Observed 10log Bayes Factor
rs2979489
RBPMS
GTF2E2
GSR
30367
30500
30633
Chromosome 8 position (kb)


MCH; finemapping RBPMS locus; MANTRA meta-analysis; MultiEthnic
Observed 10log Bayes Factor
rs2979489
RBPMS
GTF2E2
GSR
30367
30500
30633
Chromosome 8 position (kb)


MCV; finemapping RBPMS locus; MANTRA meta-analysis; MultiEthnic
Observed 10log Bayes Factor
rs2979489
RBPMS
GTF2E2
GSR
30367
30500
30633
Chromosome 8 position (kb)


287 kb
RBPMS
rs2979489
rs2979489
GATA2
GATA1
ATAC
H3K27ac
D0
H6
D3
D4
D5
RBPMS
46 KB
46 KB


be3 globin
at 16-18 ss
kdrl
at 24-26 hpf
benzidine staining
at 48 hpf
uninjected
gtf2e2 MO
4 ng
rbpms2a MO
6 ng
rbpms2b MO
4 ng
rbpms MO
1ng
A
A'
A''
57/60
29/32
42/43
B
B'
B''
18/22
29/31
18/22
C
C'
C''
33/33
13/18
30/32
D
D'
D''
19/22
13/18
17/20
E
E'
E''
9/17
28/28
67/68